ISP Project Report

Ivan Fursov

1 February 2019

1 Introduction

The problem most insurance companies face is fraud which causes large financial losses. Fraud detection techniques can help to automate the process of investigating dishonest companies.

The goal of this project was to build a deep learning system that detects medical fraudulent receipts and to estimate how well embedding approach work on non-textual data, particularly on the data provided by Allianz insurance company.

2 Data

The dataset consists of 4 files with 3.2 Mln samples in total. There is 1.5% of fraudulent receipts. Each sample represents one medical prescription for a particular patient. The following features are known for each row:

- ID — Unique id
- KORREKTUR — Adjustment amount
- RECHNUNGSBETRAG — Invoice amount
- ALTER — Customer age
- GESCHLECHT — Gender
- VERSICHERUNG — Insurance type

- FACHRICHTUNG — Doctor's specialty
- NUMMER — Treatment
- NUMMER KAT — Treatment group
- ANZAHL — Number of treatments
- FAKTOR — Increase factor
- BETRAG — Treatment cost
- ART — Material cost
- TYP — Billing type
- LEISTUNG — Benefits type

The target column was created as $y_i = 1$ if $KORREKTUR > 0$ and $y_i = 0$ otherwise.

3 Experiments

Fraud detection can be viewed as a binary classification problem. Assume a set of receipts R, where each receipt $r_{ij} \in R$ is associated with a list of sequence of treatments for a particular patient $\{p_1, p_2, \cdots, p_m\}$, together with their labels $\{y_1, y_2, \cdots, y_m\}$, where $y_i = 1$ if the sequence of treatments contains a fraudulent receipts and $y_i = 0$ otherwise. Thus, the task is to learn a classifier over sequence of treatments so that it can predict the fraudulent receipts.

3.1 Models

For each model there were two types of features available. The first is sequence of treatments. The second is meta-information about the client: gender, age, insurance type and invoice amount. To represent treatments we used three approaches:

Bag-of-Words: this approach represents a sequence of treatments as a sparse vector with 1-s one the i-th position of t_i treatment and 0-s elsewhere.

TF-IDF: this is a weighted BoW approach. The intuition behind it is that if a treatment occurs multiple times in a document, its relevance is increased. At the same time, if a treatment occurs many times in many other patients' receipts, its relevance should be lowered.

Embeddings: Each treatment t_i is represented as a vector v_i of dimension d and computed by embedding each t_i in a continuous space, using an embedding matrix E (of size $d \times |V|$). Thus, the sequence of treatments $t_1, t_2, \cdots, t_m$ is embedded into T matrix of size $d \times m$.

3.1.1 XGBoost

The first classifier is **XGBClassifier** from **xgboost** package. We used it as a baseline solution for the problem.

BoW and Tf-Idf representations of treatments along with additional features were used as inputs to XGBClassifier model.

3.1.2 SWEM model

Two ways of obtaining embeddings were used: pretraining them using continous bag-of-words (CBOW) model introduced in [2] and initializing with truncated random normal distribution ($\mu = 0$, $\sigma = 0.001$) and training them end-to-end.

Given treatmnet embeddings, the bag-of-words model generates the vector representation of a sequence of treatments by averaging (1) or calculating maximum in each dimension (2) over the embeddings of all treatments. The vector is then passed through two fully connected layers with ReLU activation. Simple Word-Embedding based Models (SWEMs) are introduced in [3] and show strong performance in many NLP tasks.

To take additional features into account, we included extra tower with fully-connected layers over meta-features. The outputs of treatment tower and

feature tower are concatenated before passing through two fully connected layers with ReLU activation.

$$v = \frac{1}{|T|} \sum_{i}^{|T|} t_i \tag{1}$$

$$v = \max(T, axis = 1) \tag{2}$$

3.2 Training

The models were trained for 5 epochs using Adam optimization algorithm [1], minimizing a standard cross-entropy loss. All training uses a batch size of 2048 and learning rate of $\alpha = 0.001$. The weights were initialized randomly from a Gaussian distribution with zero mean and $\sigma = 0.01$.

Embedding matrix E was trained using **Word2Vec** model from **Gensim** package with $window\_size = 10$, $min\_count = 2$ and other hyper parameters with default settings.

3.3 Metrics

The evaluation of models was performed on the validation set. The data was split into train and test sets in the ratio of 9:1 respectively.

ROC AUC and Average PR metrics were used to evaluate the effectiveness of models.

3.4 Results

The main results across all the models are given in Table 1. All embedding-based models outperform strong machine learning algorithm.

Conclusion

This research showed the effectiveness of applying embedding approach to non-textual data in fraud detection task. It demonstrated that this approach

Model	Features	ROC AUC	Average PR
Xgboost (BoW)	w/	86.6	14.2
Xgboost (BoW)	w/o	82.2	13.1
Xgboost (Tf-Idf)	w/	86.8	13.8
Xgboost (Tf-idf)	w/o	84.3	13.4
SWEM-max (d=300, not pretrained)	w/	87.1	15.5
SWEM-max (d=300, not pretrained)	w/o	86.3	14.3
SWEM-max (d=300, pretrained)	w/	**87.5**	**15.7**
SWEM-max (d=300, pretrained)	w/o	86.9	15.2

Table 1: Results.

with a simple DL model significantly improves performance over the strong ML methods.

References

[1] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014.

[2] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781, 2013.

[3] Dinghan Shen, Guoyin Wang, Wenlin Wang, Martin Renqiang Min, Qinliang Su, Yizhe Zhang, Chunyuan Li, Ricardo Henao, and Lawrence Carin. Baseline needs more love: On simple word-embedding-based models and associated pooling mechanisms. *CoRR*, abs/1805.09843, 2018.